Empirical evidence in conceptual engineering, or the defense of 'predictive understanding'

Wiktor Rorot[1], Marcin Miłkowski[2]

April 2024

Preprint before proofreading, forthcoming in: Stalmaszczyk P. (Ed.). *Current Issues in Conceptual Engineering: Methodological and Meta-philosophical Considerations*. BRILL mentis.

Abstract

The advent of digital philosophy of science—the practice of employing large textual datasets together with text mining and natural language processing tools within philosophy of science—has dramatically shifted philosophers' ability to account for the scientific practice, without relying on small and often arbitrary samples of the literature. Instead, digital philosophy of science allows for investigating tens of thousands of academic papers, not to mention other genres of scientific communication, providing a broader perspective on the scientific practice. The methodological issues that come with this novel approach have been traditionally discussed around the necessary choice between using these methods for discovery and exploration of new philosophical hypothesis, and the testing of existing stances. In this chapter, we propose a different path that remains hitherto underexplored in digital philosophy: namely, one offered by conceptual engineering. We provide general methodological guidelines for using digital tools in the study of scientific literature for the purpose of clarifying and honing existing philosophical concepts. To substantiate this position, we draw examples from existing research which has hinted at this approach and apply our proposed methodology to the notion of "understanding" in the scientific practice to show how it can help researchers across subdisciplines of philosophy apply this concept usefully, including in the philosophy of artificial intelligence.

Keywords

Conceptual engineering; understanding; explainable AI; digital philosophy of science; prediction; explanation; text mining; digital humanities;

[1] University of Warsaw, w.rorot@uw.edu.pl

[2] Institute of Philosophy and Sociology, Polish Academy of Sciences, marcin.milkowski@ifispan.edu.pl

1 Introduction

The twentieth century witnessed a sudden surge in the popularity of the philosophy of science. Notably, in the aftermath of the Second World War, precisely between the years 1940 and 1960, the frequency of the term's usage increased by over 3.5 times, as ascertained through a search of the Google Books Ngram corpus.

This escalating appeal can likely be attributed to the concurrent expansion of the scientific enterprise itself, as suggested by Ankeny (see Ankeny et al., 2011), an influence that may have significantly shaped the trajectory of philosophy of science in the twenty-first century. Notably, this evolution has been characterized by a growing departure from traditional "armchair" philosophy, favoring a heightened reliance on various sources of empirical data. Broadly characterized as a methodology reminiscent of that employed in the natural sciences, this reliance manifests in two primary forms: naturalistic or empirical philosophy of science, which employs empirical methods like those drawn from the history of science, as well as social sciences in the philosophical analysis of science, and the philosophy of science in practice, which introduced a focus on science as a real, human enterprise.

Digital philosophy of science is an approach that has grown at the intersection of these two frameworks. Drawing upon the techniques of natural language processing and the methodologies of digital humanities, it capitalizes on the advantages offered by computational tools. These tools empower researchers to explore a vast volume of data that was previously beyond the limits of human cognitive capabilities. Digital philosophy of science in particular focuses on the study of large corpora of scientific communication, to uncover general tendencies in how scientists approach epistemic and methodological issues.

Throughout its evolution, empirical or naturalistic philosophy of science has primarily relied on in-depth case studies of scientific practices. These studies, however, inherently possess certain limitations. They often exhibit a narrow focus, and many philosophical discussions have centered on the same set of examples, as scholars debated interpretations presented in existing literature (Pitt, 2001). While empirical philosophy of science approach undeniably yielded valuable insights into the inner workings of science, it was encumbered by significant constraints. The selection of examples was often confined to those already familiar to the philosophers, thus restricting their ability to generalize the conclusions drawn from these studies (see e.g., Currie 2015). It is precisely this limitation that has spurred the advancement and application of digital tools for the analysis of scientific papers.

A similar dynamic is currently at play in the realm of conceptual engineering. The methods of conceptual engineers, with their focus on fine details, necessitate concentration on a limited set of texts and arguments while relying on external evidence, unrelated to their practice, to establish the philosophical significance of this chosen sample. This, however, does not guarantee that the selected sample is representative of the broader utilization of the concept in question. As an example, consider the conceptual engineering of the concept of a "set" as advanced by Fenner Stanley Tanswell (2017). In order to evaluate and improve the concept, Tanswell analyzed two papers in detail, without the

possibility to say how the concept is employed across a larger body of studies in set theory, or across the applications of set theory outside of logic. For this reason, we propose in this chapter that the methodology of digital philosophy (of science) can prove invaluable in expanding and fortifying the endeavors of conceptual engineers.

To offer a way of employing the methodology of digital philosophy in conceptual engineering, we introduce several essential practical aspects of the digital tools that will be referenced in this chapter. This preliminary step is the focus of Section 2 and is intentionally concise, with additional references provided for readers who wish to delve deeper into the subject matter. Section 3 lays out our principal theme, specifically, the vital role that empirical evidence can play within the conceptual engineering enterprise, and conversely, how the perspective of conceptual engineering can help resolve the tension surrounding the utilization of digital methods for either the discovery or testing of hypotheses. Section 4 serves to exemplify the preceding discussions with a concrete case from our own research—an examination of the concept of "understanding" in scientific practice. Digital methods uncover the relevance of prediction, which has been overlooked in philosophical accounts of "understanding". The impact of our revised concept of "predictive understanding" is the most visible in the domain of philosophy of artificial intelligence. We explore in detail these practical insights, showing how noting the role of prediction in understanding impacts policy-making for AI technologies. More broadly, this example shows the relevance of conceptual engineering.

2 Digital tools for philosophy

2.1 Citation analysis and text mining

In the study of scientific literature within the realm of philosophy of science, two primary categories of tools have gained widespread adoption: citation analysis and text mining. Citation analysis enables the construction and examination of citation networks, which involves the scrutiny of relationships between various publications, exploration of the topology of scientific fields, and occasionally incorporates a more intricate examination of the context in which a particular paper emerged. This approach has garnered significant interest within the field of social science of science. Philosophers have also harnessed this method, utilizing it, for instance, to investigate the epistemological evolution of science (see Chavalarias & Cointet, 2013). Nevertheless, this approach is only peripherally related to conceptual engineering. Consequently, in this section, our focus shifts to the methods of text mining.

2.2 Text mining

2.2.1 Background: Distributional semantics

An early linguistic observation that has shaped the development of modeling semantics in large language corpora has been formulated by John Firth (1957, p. 11): "You shall know a word by the company it keeps!" Firth discusses the role that collocations, that is patterns of co-occurrence of words, play in linguistic research and specifically lexicography. As his views are associated with a Wittgensteinian perspective on language (Wittgenstein, 1968), this claim indicates not only the

methodological recommendation regarding the study of semantics but also has a stronger sense: for Firth, at least part of the meaning of a word is, in fact, established by its frequent neighbors. This approach has been, in fact, recently dominant in the Natural Language Processing (NLP) research community's attempts to provide quantification of semantics of natural language (see Mitchell, 2019, ch. 11; Lenci, 2008, 2018), commonly known as distributional semantics.

Distributional semantics provides not only a particular representation of meaning in terms of a vector of frequency distributions, but also a computational toolkit for deriving these representations from empirical data (Lenci, 2018, p. 152), primarily sourced from the extensive textual corpora available today. Within this framework, words are depicted as vectors that capture their co-occurrences with other words in the linguistic contexts found in the corpus. These vector representations are generated through a process that seeks to learn the probabilities of these co-occurrences while simultaneously reducing the dimensionality of this representation, ensuring both computational tractability and generalizability of the algorithm.

This approach furnishes a geometric representation of words within the corpus, enabling various semi-quantitative analyses of their meanings. Researchers can explore the proximity of words in this geometric space, often interpreted as an indicator of their semantic relatedness, or perform mathematical operations on these vectors to investigate semantic analogies. For instance, this method can assist in identifying words that share similar semantic relationships, as in the analogy "word" is to "language" as "note" is to "music" (see Turney, 2013). The core principle of this algorithm involves comparing the discrepancies between pairs of vectors.

This approach has played a pivotal role in driving the contemporary success of large language models such as GPT (OpenAI, 2023) and BERT Mickus et al. (2020). Furthermore, it has been explored as a potential model of semantics in the human brain (see, for instance, Dove, 2014) and as a means to address the philosophical challenge of naturalizing meaning and incorporating it within the framework of Shannon information theory (Isaac, 2019).

2.2.2 Text mining – an introduction

The digital philosophy of science is built upon the foundation of digital humanities, which harnesses this fundamental approach in a variety of ways. Among these, two methods have gained prominence: firstly, the "simple" co-occurrence analysis, which will be the focus of this chapter, and secondly, the topic modeling methodology (a good overview of this approach, aimed at philosophical audiences, is offered by Lean et al., 2021; Malaterre et al., 2019).

Frequency and co-occurrence analysis

A qualitative exploration of frequencies of central concepts and notable co-occurrences, as identified by the tools of distributional semantics, was employed in one of the earliest digital studies in the field of philosophy of science. In a paper summarizing his doctoral dissertation, James Overton (2013), delved into the role of explanation in scientific practice. Initially, Overton compiled a list of theoretically relevant terms associated with various aims of science, as discussed in the philosophical

literature. Subsequently, he conducted comprehensive searches within a corpus encompassing all articles published in the journal *Science* over the course of one year.

This approach furnished him with empirical evidence concerning the significance of explanations within scientific practice, with terms related to "explain" appearing in nearly half of his sample, more frequently than words pertaining to "theory" and others. Following a thorough examination of a random subsample, Overton identified five distinct categories of explanations offered by scientists. He then proceeded to investigate how existing philosophical perspectives fared in accounting for these various categories.

Outside of philosophy of science, co-occurrence analysis has also found application in the work of Mark Alfano (2019). Focusing on Nietzsche's moral psychology, Alfano sought to move beyond the examination of seemingly unverified and superficial connections among the various terms Nietzsche utilized, which, in his assessment, had dominated the secondary literature on this subject. To achieve this objective, he embarked on a diachronic study of Nietzsche's body of work, systematically investigating the co-occurrences of terms throughout his diverse writings. This investigation led to the construction of a conceptual network representing Nietzsche's moral psychology and underscored the changes it had undergone over the course of his career.

Alfano's approach unveiled that certain concepts which had been the primary focus of Nietzsche's readers, such as "resentment" and "will to power," were infrequent within his oeuvre and remained disconnected from the central network of concepts he employed, as indicated by the paucity of their co-occurrences in close textual proximity.

What is common to both of these distinct topics and methodologies is that digital tools are employed to unveil usage patterns of relevant concepts that would otherwise remain unnoticed during a close reading of a limited selection of texts. Furthermore, these tools demonstrate that, while "simple" co-occurrence alone may not provide a straightforward definition of a concept, it does illuminate its critical semantic attributes, which any comprehensive definition should encompass. This significance becomes particularly evident in the application of this methodology within contemporary lexicography (Horák & Rambousek, 2018), where the practice is enriched by these methods and extended, for instance, by incorporating information about the syntactic structure of co-occurrences within distributional models (see Kilgarriff et al., 2014).

Multi-level methodologies of digital studies

The methodologies of text mining underscore a crucial element of the digital approach: while it can provide robust, empirically grounded data for the philosopher, it must be complemented by meticulous qualitative analysis and interpretative work (see Murdock et al., 2017). The prevalent methodological framework, widely accepted, though sometimes implicit, involves a twofold investigation. In the initial step, digital tools are employed for what is often referred to as "distant reading" (Moretti, 2013). This step allows for the extraction of high-level patterns that are discernible only at the level of the broader knowledge system and accessible solely through the use of computational tools.

However, for these results to serve as valuable resources for philosophers, they must then undergo a second phase of interpretative close reading. This phase encompasses not only the results themselves, typically presented in the form of visualizations, but also an examination of the texts from which they were extracted. It involves identifying relevant subsamples of the corpus that are amenable to close reading (i.e., of a manageable size) and are chosen based on objective criteria, as opposed to arbitrary selection, which is often seen in conventional case studies within the philosophy of science. These subsamples are shown in the first step to have broader relevance within the context of the philosophical inquiries at hand. One of the key benefits of this procedure is that it allows a human validation of the results, a step that is seen as a desired practice in digital humanities (see Pichler & Reiter, 2022).

3 Digital empirical evidence for conceptual engineering

Conceptual engineering, a philosophical methodology that aims to revise concepts–variously understood as psychological concepts, words' intensions, commitments and social norms, and others (see Koch et al. 2023),– possesses features that align it particularly well with the digital methods described earlier. Conventionally, conceptual engineering is juxtaposed with conceptual analysis, which is primarily concerned with the description and formalization of concepts. In contrast, the mission of conceptual engineers is one of *explication* (Carnap, 1967): to transform a vague or imprecise concept, referred to as the *explicandum*, into a new, precise concept, referred to as the *explicatum*. This process necessitates an initial phase in which the concept is critically examined, and any deficiencies are identified (Cappelen, 2018). Cappelen delineates two primary categories of deficiencies (Cappelen, 2018, p. 34): the concept might lack semantic value, rendering it incoherent or nonsensical, or the attributed semantic value of a particular expression may be in some way detrimental. Although extensive efforts have been dedicated to identifying morally, politically, or socially problematic concepts (the most famous example is the work of Sally Haslanger, e.g., Haslanger 2003), and the methods we discuss can be applied in these instances. Our motivation here is that there may be adverse effects of unclear or underdeveloped concepts on theorizing and the broader scientific practice.

Identifying deficiencies in concepts can be approached from various angles, and one significant tradition in philosophy involves employing thought experiments for this purpose. Consider, for instance, the role of Gettier cases in reevaluating the concept of "knowledge" (Gettier, 1963; Sękowski, 2022), or the examples of cognitive extension presented by Clark and Chalmers to challenge the adequacy of the concept of "belief" (Clark & Chalmers, 1998). Nevertheless, the escalating skepticism surrounding philosophers' intuitions (an early critique of this aspect is found in Machery, 2004) and "armchair" methodologies has prompted a shift towards incorporating empirical evidence. This is manifest through the emergence of fields like experimental philosophy (refer to, for instance, Stich & Tobia, 2016), or the integration of pertinent findings from psychological and cognitive science research (see, for example, Leslie, 2017). In this context, we propose that digital data can be similarly employed.

In the realm of conceptual engineering, the tools offered by contemporary computational linguistics provide the means to draw inferences regarding the semantics of words and concepts from a vast and varied corpus of language data. These capabilities represent a pivotal shift away from the reliance on potentially biased and limited samples of word or concept usage that researchers may have encountered themselves. This transition holds substantial importance when assessing the deficits in concepts. Analogous to how the "negative program" of experimental philosophy (Stich & Tobia, 2016) illuminated the inherent biases in philosophers' intuitions, leveraging large language corpora to explore concept semantics empowers us to unearth and substantiate claims regarding concept usage and its real-world implications.

Using computational linguistics methodologies facilitates the discovery of concepts that may have eluded prior philosophical inquiries, as exemplified by the concept of "predictive understanding," which will be examined in the following section. These methodologies can further be expanded to enable detailed analyses of the interrelationships between various concepts, thereby forging new paths for identifying conceptual inconsistencies and probing into the social or moral repercussions of these concepts. For now, we will primarily focus on the discovery of concepts.

The digital methods hold particular value for subdisciplines of philosophy that strive to bridge technical philosophical notions with non-philosophical terms, as is the case for philosophy of science. In this context, the application of digital empirical evidence enables a robust bottom-up exploration strategy, guiding philosophical inquiry with concrete evidence. The approach unfolds as follows: once we pinpoint our primary area of interest (for instance, the concept of "understanding," the central case study in this chapter), we proceed with the digital analysis of available data, as previously discussed. The aim of this analysis is to unearth and categorize usage patterns that align with the philosophical understanding of key concepts, as applied in non-philosophical contexts, such as scientific practices and their textual outputs. During such investigations, it may become evident (as we will illustrate with the notion of "predictive understanding") that certain uses do not neatly fit into previously established philosophical categories for the term. These new concepts, arising organically from the data, become valuable additions to philosophical reflections, fostering a more comprehensive and profound examination of these practices.

However, it's crucial to acknowledge certain limitations inherent in this practice. The discovery of concepts using digital methods is not entirely *a priori*; it necessitates an initial understanding of where to focus one's attention, identifying terms that might not align well with existing philosophical accounts. This human-driven identification of key areas is in fact a crucial step for any digital practice in humanities (Pichler & Reiter, 2022).

Nevertheless, after pinpointing these primary areas of interest, the extent of discovery is largely determined by the quality and scale of our data corpus. A better and larger corpus increases the likelihood of uncovering concepts that may be entirely novel in philosophical discourse. Size plays a pivotal role here given that languages are not adequately characterized by ergodic and stationary statistics, implying that certain relevant terms may only appear in corpora infrequently, if at all (see Dębowski, 2021).

Importantly, this implies that when employing digital methods for conceptual engineering, as long as we refrain from making generalizations about the significance of specific concepts within the practices under study, the risk of accepting false positives is minimal. Even if certain terms appear only rarely, the fact that they are indeed employed in the scientific practice indicates that they should be accounted for by our philosophical theories. A false positive in this case would be a term that appears in our textual evidence but bears no conceptual reality. But if such term is not nonsensical or incoherent, our philosophical theory should be capable of dealing with it—even if by explaining it away with some other concepts. In this way, even such a "false positive" offers an advancement of our philosophical account. The size of the corpus primarily helps with controlling for false negatives—i.e., increasing the likelihood that no relevant term will be missed by our analysis—larger corpora are more likely to reveal less common conceptual relationships. However, these corpora should be also balanced and representative so as to remain appropriately diverse, when quantitative analyses are performed. For example, even huge corpora of article publications may underrepresent more colloquial locutions that are typical of academic language.

This highlights the significant progress offered by digital methods. They not only enable us to transcend well-documented biases but also provide a more accurate representation of our conceptual systems as conveyed through language.

4 Engineering "understanding"

The concept of understanding holds a prominent place within contemporary epistemology and philosophy of science. It has sparked a plethora of debates that aim to elucidate various aspects of understanding, including its ontological underpinnings, the psychological processes involved, its normative dimensions, its role within the scientific enterprise, and its relationship with other key epistemic concepts such as knowledge and explanation. In recent years, the significance of this notion has grown even more pronounced.

This amplified significance is notably apparent in the domain of artificial intelligence (AI) research and AI ethics. As AI systems become increasingly complex and integral to our lives, the question of "explainability" emerges as a central point of contention. Researchers and ethicists grapple with the challenge of defining and determining what constitutes "explainability" within AI systems.

Here, we wish to argue that, despite the careful consideration within these debates, they seem to overlook a critical aspect of the concept of understanding—its relationship with prediction. "Predictive understanding" is a concept that recurrently emerges in the discourse of scientific practitioners. Surprisingly, this concept has eluded existing classifications and accounts of "understanding," even if prediction is sometimes included, along with explanation, as one of the features of scientific explanation (Regt, 2017).

4.1 Understanding "understanding"

One starting point for this inquiry can be found in the tradition of conceptual analysis of "understanding," which is influenced by the Lviv-Warsaw School, particularly the proposals of Izydora Dąmbska and Jacek Jadacki (Dąmbska, 1975; Jadacki, 1990). The Lviv-Warsaw School was an influential philosophical movement, founded by Kazimierz Twardowski in the late 19th century, that made significant contributions to logic, analytic philosophy, and related fields through the work of members such as Kazimierz Ajdukiewicz, Tadeusz Kotarbiński, Janina Kotarbińska, Stanisław Leśniewski, Jan Łukasiewicz, Alfred Tarski, and Izydora Dąmbska. The common idea running through proposals of Dąmbska and Jadacki is that "understanding" is fundamentally a polysemous term. In their efforts to disentangle the various concepts hidden within this term, Dąmbska and Jadacki seek to identify a broader natural kind to which these concepts may belong.

Dąmbska's approach aligns with what Kim (1994) terms Hempelian *explanatory internalism*. As Kim puts it, "whether or not something is explanation–a good, 'true' or 'correct' explanation–depends on factors internal to a body of knowledge, not what goes on in the world" (p. 57). Similarly, for Dąmbska, understanding is a relation that takes place between mental states, such as beliefs or propositions. Additionally, Dąmbska introduces a pragmatic dimension to her proposal, stipulating that genuine understanding involves the capacity to recall or apply the object of understanding in different contexts (Dąmbska, 1975, p. 53).

On the other hand, Jadacki, in his analysis of ordinary Polish[3], identifies four distinct senses of the verb "rozumieć" (to understand), as follows:

1. to grasp intuitively ["intuitive understanding"]
2. to be aware of (know) ["identificatory understanding"],
3. to accept (justify) ["indulgent understanding"],
4. signify (mean) ["inscriptive understanding"].

While scientific discussions may occasionally invoke the fourth sense, especially when expressing concerns about the intelligibility of terminology or arguments, this aspect is not our focus in this paper. Similarly, the moral or approbatory sense of "understand" (sense 3) is not pertinent to the philosophical discourse on scientific understanding, just as intuitive grasp (sense 1) is not. In the realm of science, the quest for understanding transcends mere hunches and feelings. Consequently, the only sense of "understanding" relevant to the philosophical debate is the second sense.

---

[3] While Jadacki's analysis primarily centers on the Polish language, he points out (1990, p. 18) that these senses are not confined to Polish alone. Instead, they can be discerned in numerous other languages, spanning Russian, English, German, Italian, French, as well as ancient Latin and Greek.

In his analysis, Jadacki distinguishes various subtypes of identificatory understanding, highlighting that one can grasp different aspects of X:

- Its (essential) nature.
- Its structure.
- Its context (or situation).
- Its causes.
- Its effects.
- Its rationales (or reasons).
- Its category (or type).

It's noteworthy that all these features can be provided as responses to various knowledge-seeking inquiries, particularly when framed in terms of wh-questions: What is it? How is it structured? When is it found? Why does it exist? What does it cause? What is its purpose? What category does it belong to? This typology serves as a comprehensive framework. However, it's important to recognize that not all answers necessarily entail knowledge in terms of causal explanations. For instance, understanding the structure of a musical composition allows us to appreciate the piece without having a causal or mechanistic understanding of its creation. In other words, Jadacki's analysis appears to encompass both descriptive and explanatory factors that come into play when we engage in understanding in the sense of identification.

This perspective can be contrasted with the stance advocated by Michael Strevens (2013), who also distinguishes between "understanding that" and "understanding why," with the latter being the relevant concept in scientific understanding. He posits that "understanding why" entails "grasping a correct explanation" and proceeds to elaborate and defend this "simple" view. While this perspective retains the pragmatic aspect of understanding found in Dąmbska's proposal and recognizes the polysemous nature, akin to Jadacki's analysis, it tightly associates understanding with explanation, disregarding the possibility that description and prediction could be integral components of scientific understanding. Similarly, Kareem Khalifa (2017) claims that proper understanding is provided only by correct explanations. This perspective has maintained a presence in the philosophy of science (see Elgin, 2007; Potochnik, 2017). Our position is closer to the one defended recently by Dellsén (2016), who emphasizes that a *complete* understanding of a given target would require one to grasp how to correctly explain and predict every aspect of the target. However, we believe that identificatory understanding in science involves grasping more aspects of the target, making us more pluralistic. Interesingly, our pluralism is entirely consistent with Wilkenfield's (2019) view, who claims that attaining more understanding consists in the ability to generate more useful information from an accurate, more minimal representation. Simply, we believe there are multiple kinds of such information.

4.2 Digital investigation of "understanding"

With this clearly delineated focus, we can now initiate our examination of the specific sense of the term "understanding." To achieve this, we will employ digital tools on the corpus of publicly available scientific paper peer reviews. Specifically, we will analyze a dataset published by our research group: the corpus containing 22,819 open peer reviews for papers from the *eLife* journal (Miłkowski & Jasieński, 2022). Our choice of this dataset is motivated by two key factors: the first is that we want to cover the academic English as used both in publications and in less formal genres, such as peer reviews. The second reason relates to the reviews' purpose, which entails, at least in part, the appraisal of a paper's epistemic significance. Therefore, we anticipate that a higher number of epistemic terms will be present in these reviews compared to standard scientific discourse, where epistemic concepts are employed more implicitly, thus enhancing the robustness of our study.

4.2.1 Methods & results

In the initial step, we identified 408 documents containing the term 'understanding' [4]. These documents were subsequently uploaded to the web-based tool SketchEngine (Kilgarriff et al., 2014).[5] Before delving into the analysis of our corpus, it is advantageous to present a visual representation of synonyms and related terms for "understanding" as they appear in scientific discourse broadly. To achieve this, we will rely on the corpus of the Directory of Open Access Journals (DOAJ) as made available through SketchEngine, which encompasses 2.7 billion words from open-access journal papers spanning from the 19th century until the 2010s. The thesaurus function of SketchEngine has generated FIGURE 2, depicting terms that share similar meanings with "understanding." Notably, we observe the presence of "explanation," but we also find terms such as "description" and "assessment."

[4] Please note that this study is intended primarily as a proof-of-concept and should eventually be replicated on a larger and more diverse corpus of reviews that is currently under construction. Technical details can be found in the Supplementary notes.

[5] The SketchEngine tool is available at sketchengine.eu.

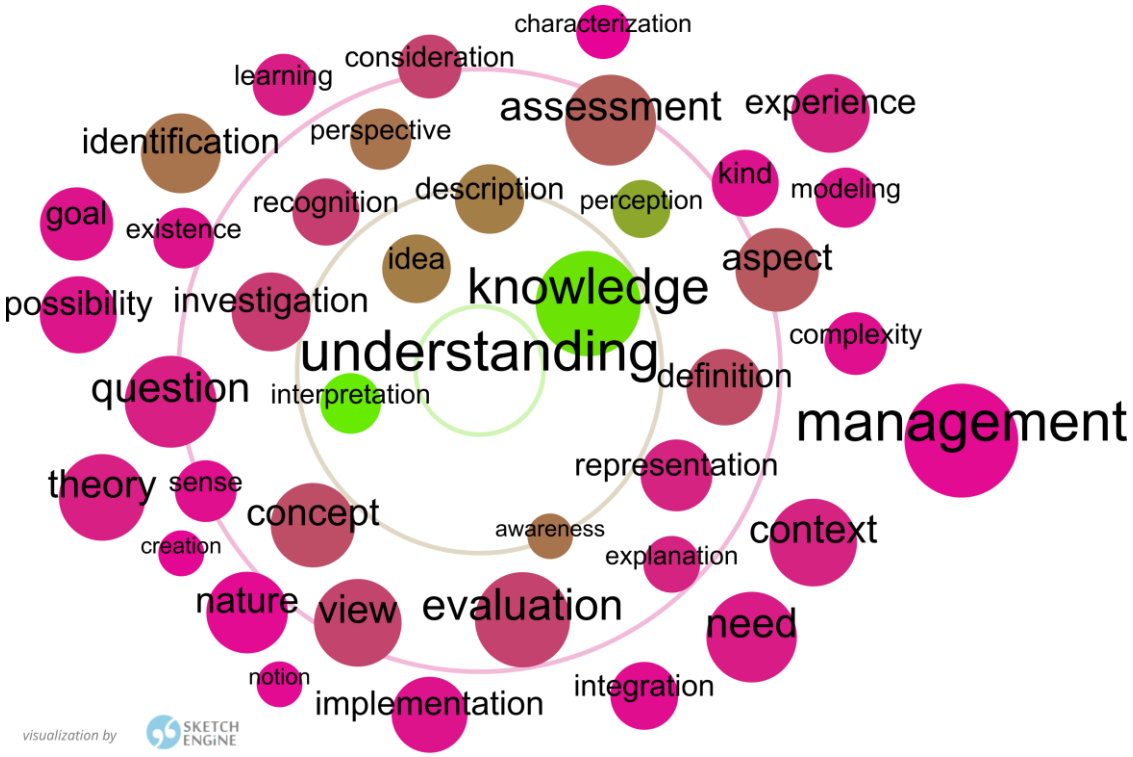

Figure 2: Near-Synonymous Terms for 'Understanding' in the DOAJ Corpus.

The DOAJ corpus contains 115 examples of 'predictive understanding.' However, we decided to extend our analysis beyond the realm of publications. Hence, we turned to our eLife open peer review corpus, which includes both articles, referee reports and author responses. Leveraging SketchEngine, we extracted "word sketches," which are distributional representations of word collocations that reflect the surface grammar of language. These word sketches were scrutinized with a primary emphasis on the modifiers employed in conjunction with the target noun, as illustrated in Figure 3. This focus on modifiers is essential since they likely correspond to the various forms of understanding as manifested in academic discourse.

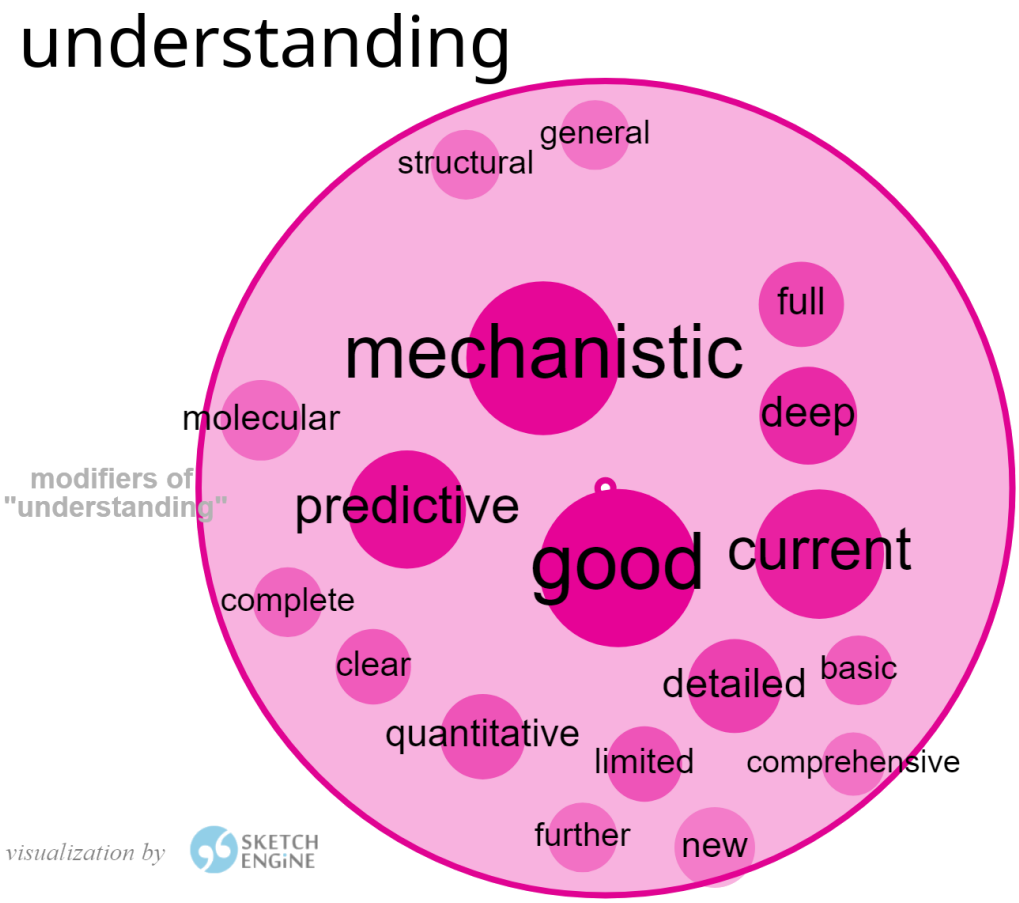

Figure 3: Modifiers of "understanding" in the eLife open peer review corpus.

The most typical modifier is "good," as reviewers frequently comment that some revisions are needed "for a better understanding." Similarly, modifiers like "deep," "detailed," or "full" often pertain to whether a reader can gain a full understanding or deep understanding of the issues described in the paper under review, or whether the paper provides all the necessary details and depth, among other factors. Additionally, these modifiers are used to indicate the current state of scientific knowledge when referring to "current understanding."

However, there are several modifiers that appear relevant to philosophical debates about scientific understanding. These include terms like "molecular," "mechanistic," "structural," "quantitative," and, notably, "predictive." It is interesting to observe that no corpus of open reviews we compiled contains the term "explanatory understanding," which remains central to philosophical discussions[6]. Instead,

[6] It may be due to the fact that Strevens' simple view is broadly and implicitly accepted, meaning that "explanatory understanding" serves as the overarching term for all forms of scientific understanding - i.e., any unqualified appearance of "understanding" refers to "explanatory understanding". We find this explanation, nonetheless, unsatisfactory and underdeveloped. Further research would be needed to motivate such perspective more substantially.

reviewers and authors often mention “mechanistic understanding,” which appears to be related to the mechanistic account of explanation (Machamer et al., 2000).

In our corpus 16 documents refer to “predictive understanding” (including 5 repetitions due to five mentions of a paper title: “Challenges in microbial ecology: building predictive understanding of community function and dynamics”). While these numbers may not be overwhelming, they reflect a consistent pattern of usage, aligning with the findings from the DOAJ corpus. This consistency encourages us to reconsider whether various types of understanding correspond to the general aims of science, which have traditionally been analyzed in terms of description, prediction, explanation, and control. Although this list of aims in science may be incomplete (Miłkowski, 2022) and the notion of aims of science has been critiqued (see Rowbottom, 2014), it is not entirely without merit. Identificatory understanding can be achieved through description (e.g., classifying symptoms of mental disorders), explanation (e.g., mechanistically), as well as through prediction. These are distinct types of understanding that have been largely overlooked in the philosophical literature. In this paper, we will set aside the issue of understanding via control, as it is typically discussed under the broader concept of know-how or “phronesis” and cannot be readily identified in our sample.

The quantitative perspective on predictive understanding is most effectively complemented by a qualitative examination of corpus concordances. Here are some typical usages:

1. Only with the combined ability to assay synthetic and endogenous gene circuits can we develop a predictive understanding of the gene regulatory code underlying cellular decision programs. (Anonymous Reviewer & Teichmann, n.d.)

2. Extending this framework within and beyond biomechanics may yield a predictive understanding of the tempo and mode of evolutionary diversification. (Muñoz et al., 2018)

3. A more complete analysis of the cis- and trans- interactions among all ligand-receptor pairs, for different levels of Fringe expression, could provide a more comprehensive and predictive understanding of how cells with distinct component combinations signal to their neighbors, and to themselves, through Notch. (Nandagopal et al., 2019)

4. Achieving a quantitative and predictive understanding of 3D genome architecture remains a major challenge, as it requires quantitative measurements of the key proteins involved. (Cattoglio et al., 2019)

5. However, in most cases we remain far from a predictive understanding of how the specific evolutionary forces imposed on a given interacting pair of proteins at the molecular level translate into a given process of functional diversification. (Chantreau et al., 2019)

6. Thus, our theory-experiment dialogue uncovered potential molecular mechanisms of transcriptional regulatory dynamics, a key step toward reaching a predictive understanding of developmental decision-making. (Eck et al., 2020)

In cases 1, 2, 5, and 6, the term "predictive understanding" essentially equates to "prediction" in a more complex and academic phrasing. It is worth considering whether the term could be replaced with "prediction" for the sake of clarity, as academic English can sometimes be overly verbose. However, in case 1, "predictive understanding" is set in contrast to resolving the composition of the human gut microbiome, and grasping the structures is one of the subtypes of identificatory understanding discussed by Jadacki. This goes beyond mere prediction and suggests a need for a prediction that establishes a connection between structures and their development. Notably, developmental aspects on both ontogenetic and phylogenetic timescales are mentioned in cases 2, 5, and 6. In case 5, it appears that specific predictions are lacking to gain insights into particular processes or mechanisms. Therefore, we can observe functional diversification without the ability to fully predict the evolutionary forces that trigger it in a specific manner. This implies that we may need more than just a retrospective explanation of a given process or mechanism; we require detailed predictive insights into its causal structure, which cannot be reduced to explanatory insights. It's well-established that we may lack predictions for systems that can be readily explained by conforming to a law, such as the three-body problem in classical mechanics. However, not all predictions, especially those generated by "brute force," contribute to scientific understanding. Predictions hold value when they offer principled insights.

Now, let's examine cases 3 and 4, where "predictive understanding" is further qualified by adjectives such as "comprehensive" and "quantitative." "Comprehensive understanding," a hallmark of scientific systematicity (Rescher, 1979), is evidently more desirable than understanding that is patchy and incomplete. Likewise, quantifiability equips us with a mathematical grasp of reality, enhancing our capacity for deeper understanding.

In essence, these cases indicate the existence of a distinct form of understanding that science can furnish—one that extends beyond being purely descriptive (structural), causal (or mechanistic), or exclusively explanatory. Instead, it encompasses the realm of prediction.

4.2.2 Using digital evidence for conceptual engineering

The observation that understanding in science involves prediction, however, does not constitute a conclusive argument for a conceptual engineer. A mere descriptive account of the present usage of the rather complex expression "predictive understanding," which appears noteworthy but not exceedingly prevalent in scientific literature, falls short of justifying the proposition that a philosophical concept should be redefined. One possible response to our analysis is that this expression is yet another example of the typical verbosity often found in academic discourse.

To counter this objection, we must offer a rationale for the introduction of this new expression. The situation seems to be as follows:

- "Predictive understanding" conveys more than just "prediction." The latter can be achieved without a deep comprehension of the underlying phenomena, as demonstrated by inscrutable machine learning models applied to various phenomena.

- Peer reviews typically belong to the tersest forms of writing, often composed hastily. On average, they encompass 931 words (with a median of 536) in our most extensive corpus, the MDPI Open Peer Review Corpus 2 (v1) (Miłkowski et al., 2023). In contrast, in our eLife open peer review corpus, the average word count is notably higher at 2063.7, with a median value of 1549: still much more concise than scientific papers.

This rationale leads us to believe that verbosity is less likely to manifest in peer reviews than in scientific publications. However, a definitive conclusion regarding this hypothesis would require a more extensive and focused investigation, a task that we leave for future endeavors. Furthermore, given the resemblance of "predictive understanding" to other forms of understanding, particularly mechanistic, causal, explanatory, and molecular understanding, we find it valuable to incorporate this term into the philosophical lexicon.

Additionally, it's worth contemplating the distinct value of prediction, which differs from that of explanation. In the philosophy of science, it is well-acknowledged that the symmetry thesis, as advocated by Hempel and Oppenheim (1948) in their classical account of covering law explanation, cannot be maintained. According to their model, both predictions and explanations share the same logical structure, with the primary distinction being chronological (future events for predictions and past events for explanations). However, there exist predictions without accompanying explanations (such as predicting the length of a pole by observing its shadow, which is not explanatory), as well as explanations without predictions (as seen in the case of the three-body problem).

Prediction, though sometimes undervalued in contemporary philosophy of science, is a fundamental objective of many scientific undertakings (Yarkoni & Westfall, 2017; see also Douglas, 2009). This justifies, at least from our perspective, the introduction of the term "predictive understanding" into the broader discourse surrounding understanding.

4.3 Predictive understanding for explainable AI

The predictive aspect of understanding is particularly conspicuous in the context of AI explainability. 'Explainability' in this context may mean various things: interpretability, comprehensibility, or explainability proper (Doran et al., 2017). In contrast to opaque AI systems, interpretable systems feature algorithms that can be mathematically analyzed. Comprehensible systems emit symbols enabling user-driven explanations of how a conclusion is reached. However, these need not to be factually true, thus, large language models, while comprehensible, may produce factually misleading descriptions of their decision processes. Finally, Doran et al. single out truly explainable systems, where automated reasoning is central to output crafted explanations without requiring human post-processing as a final step of the generative process. Unfortunately, the practice of explainable AI is rarely scrutinized in the light of philosophical accounts of explanation, and the notion of explanation used by practitioners remains vague and ambiguous. This makes it difficult to specify the correctness conditions for such machine-generated explanations, which also translates to the inability to provide automated benchmarks for this purpose.

While many philosophers working on explainable AI concentrate on retrospective explanation, which they link with post hoc interpretability of algorithms (Páez, 2019), the issue of explainability isn't exclusively linked to past occurrences. For instance, consider situations where an AI system denies credit or triggers an automated military response. What concerns many individuals is trust: can AI systems be relied upon to effectively address cognitive tasks, even those they haven't encountered previously? In the case of simple systems, it's sometimes feasible to establish and anticipate their future actions, providing a form of "predictive understanding" regarding their operation. The problem is that contemporary complex AI systems are not predictable in the same way, making their work inscrutable, which should also diminish our trust in their reliability.

When discussing the concept of explainable AI, we cannot solely refer to retrospective explainability, as seen in scenarios like the three-body problem, which resembles post hoc interpretability. In some respects, such retrospective interpretability is more accessible or, at the very least, more affordable than achieving a principled comprehension of how a given AI system will perform. However, the primary challenge with many contemporary, exceedingly intricate generative AI systems lies in predicting their outputs. It is crucial to be able to identify not only what is responsible for past behavior but also anticipate the future behavior of these systems, constituting two interconnected forms of identificatory understanding.

In essence, explainable AI should encompass both predictive and explanatory understanding. Drawing inspiration from practices observed in the life sciences, where predictive understanding extends beyond mere prediction, we aim to transpose these insights into the realm of explainable AI. Interestingly, de Regt (2017), in his discussion of the role of prediction in understanding, introduces the notion of an oracle producing brute predictions. He notes that:

> not only is mere prediction without understanding insufficient for science, prediction turns out to be impossible without understanding (…). Our hypothetical oracle is just a figment of the imagination: in reality one can only make successful predictions if one understands the relevant theories (Regt, 2017, p. 107).

Regrettably, de Regt overlooks the existence of such oracles in the form of opaque AI systems. Nevertheless, the imperative lies in establishing an independent method for systematically acquiring predictive and causally explanatory insights into their functioning. It is this necessity, according to our perspective, that warrants the extension of the concept of scientific understanding beyond the confines of mere explanation.

In summary, our investigation demonstrates that the term "understanding" exhibits polysemy. It also highlights that there exist forms of understanding that cannot be reduced to explanatory or descriptive understanding. In line with the philosophical tradition of the Lvov-Warsaw school, we align ourselves with those philosophers who contend that this term serves multiple functions, encompassing various types of understanding that are both attainable and valuable.

Much like explanatory pluralism has become a well-established position in discussions on explanation, we advocate for pluralism concerning understanding. While the term "understanding pluralism" may appear somewhat cumbersome, it encapsulates our perspective.

5 Conclusions

This chapter was motivated by the observation that conceptual engineering aligns with the empirical methods in philosophy of science in how it fundamentally relies on empirical evidence, particularly when evaluating the quality of concepts. While there has been a shift away from traditional "armchair" methodologies, prevailing approaches remain constrained by the cognitive limitations inherent in human information processing. To mitigate these limitations, the integration of digital tools into the practice of conceptual engineering has become increasingly relevant. This approach, often referred to as digital philosophy, aligns with the rapid advancements in computational linguistics and digital humanities, enabling philosophers to enhance their ability to extract insights from textual data without reducing their practices to mere quantitative analysis.

Here, we employ these methodologies in the context of conceptual engineering, shedding light on the concept of "predictive understanding." While philosophers of science have traditionally linked "understanding" to explanation, insights from life sciences reveal that, in certain instances, the capacity to predict specific phenomena—regardless of possessing a "correct explanation" for these predictions—can serve as a sufficient foundation for understanding these phenomena. Notably, this phenomenon is prevalent in life science, where the intricacies of complex phenomena may limit the ability to precisely model their dynamics, even with a sound mechanistic account. In such cases, the ability to predict potential interaction outcomes becomes a prerequisite for understanding these phenomena, alongside or in place of a comprehensive explanation.

Furthermore, we believe that the predictive component plays a central role in enabling us to construct genuinely 'explainable' AI systems. We strongly suspect that in the context of this phrase as used by AI system users and developers, 'explainable' should be reinterpreted as 'intelligible' or 'understandable.' This reinterpretation aligns with the vital quality we lack when dealing with highly complex deep-learning AI systems. To achieve a truly explainable AI system, we need to go beyond a profound retrospective understanding of its past performance. We should also be equipped with a principled means to predict its future behavior—specifically, its ability to consistently and accurately address problems as expected. This capability is essential for meeting users' needs by reliably and accurately solving problems. Consequently, a combination of both explanatory and predictive understanding is imperative for us to comprehend these systems.

The impact of "predictive understanding" in the study of explainable AI shows the promise of the use of digital methods in conceptual engineering, which will allow it to uncover important concepts that are rarely formulated explicitly, and can be easily overlooked with traditional approaches. Furthermore, this shows the timely role of conceptual engineering as a philosophical practice relevant to policy-making and the broader social context.

Acknowledgments

The authors would like to thank Marcin Rabiza, Richard David-Rus, Guido Löhr, and Piotr Stalmaszczyk for comments and suggestions to an earlier version of this paper. WR's work has been

supported by the scholarship provided as part of the internal grant program “Excellence Initiative - Research University” at the University of Warsaw. Licence to access the SketchEngine tool has been provided by the University of Warsaw.

Author contributions

Both authors conceived the central argument of the paper. WR wrote the first draft of sections 1, 2, 3, and 5. MM conducted the corpus study and wrote the first draft of section 4 and 5. Both authors edited the paper, prepared and accepted it for publication.

Supplementary notes

Corpus preprocessing

The *eLife* open peer review corpus was first indexed using an open-source DocFetcher software application. Subsequently, we used two queries:

1. “understanding~” – to find documents that contain this term, also in the plural form;
2. “predictive understanding” – to find documents that contain the complex term (which cannot be found in the response set due to the penalty scores used internally for document search by the Lucene engine, which implements the indexes in DocFetcher).

We also converted and filtered the contents of the corpus in JATS XML format to pure text to make sure that we did not miss any occurrences (the Java code is in our repository, available at: https://github.com/cognitive-metascience/corpora_utils).

The files found were all uploaded to SketchEngine in a single ZIP file and processed there for further analyses.

Bibliography

Alfano, M. (2019). *Nietzsche’s moral psychology*. Cambridge University Press.

Ankeny, R., Chang, H., Boumans, M., & Boon, M. (2011). Introduction: Philosophy of science in practice. *European Journal for Philosophy of Science*, *1*(3), 303–307. https://doi.org/10.1007/s13194-011-0036-4

Anonymous Reviewer, & Teichmann, S. A. (n.d.). *Decision letter: Inference of gene regulation functions from dynamic transcriptome data*. peer review. https://doi.org/10.7554/eLife.12188.034

Cappelen, H. (2018). *Fixing language: An essay on conceptual engineering* (First edition). Oxford University Press.

Carnap, R. (1967). *Logical foundations of probability*. University of Chicago Press.

Cattoglio, C., Pustova, I., Walther, N., Ho, J. J., Hantsche-Grininger, M., Inouye, C. J., Hossain, M. J., Dailey, G. M., Ellenberg, J., Darzacq, X., Tjian, R., & Hansen, A. S. (2019). Determining cellular CTCF and cohesin abundances to constrain 3D genome models. *eLife*, *8*, e40164. https://doi.org/10.7554/eLife.40164

Chantreau, M., Poux, C., Lensink, M. F., Brysbaert, G., Vekemans, X., & Castric, V. (2019). Asymmetrical diversification of the receptor-ligand interaction controlling self-incompatibility in Arabidopsis. *eLife*, *8*, e50253. https://doi.org/10.7554/eLife.50253

Chavalarias, D., & Cointet, J.-P. (2013). Phylomemetic Patterns in Science Evolution—The Rise and Fall of Scientific Fields. *PLoS ONE*, *8*(2), e54847. https://doi.org/10.1371/journal.pone.0054847

Clark, A., & Chalmers, D. (1998). The Extended Mind. *Analysis*, *58*(1), 7–19. https://doi.org/10.1093/analys/58.1.7

Currie, A. (2015). Philosophy of Science and the Curse of the Case Study. In: Daly, C. (eds) *The Palgrave Handbook of Philosophical Methods*. Palgrave Macmillan, London. https://doi.org/10.1057/9781137344557_22

Dąmbska, I. (1975). W sprawie pojęcia rozumienia. In *Znaki i myśli: wybór pism z semiotyki, teorii nauki i historii filozofii* (pp. 49–56). Państwowe Wydawnictwo Naukowe.

Dębowski, Ł. J. (2021). *Information theory meets power laws: Stochastic processes and language models*. Wiley.

Dellsén, F. (2016). Scientific progress: Knowledge versus understanding. *Studies in History and Philosophy of Science Part A*, *56*, 72–83. https://doi.org/10.1016/j.shpsa.2016.01.003

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019, May 24). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. http://arxiv.org/abs/1810.04805

Doran, D., Schulz, S., & Besold, T. R. (2017, October 2). *What Does Explainable AI Really Mean? A New Conceptualization of Perspectives*. http://arxiv.org/abs/1710.00794

Douglas, H. E. (2009). Reintroducing Prediction to Explanation. *Philosophy of Science*, *76*(4), 444–463. https://doi.org/10.1086/648111

Dove, G. (2014). Thinking in Words: Language as an Embodied Medium of Thought. *Topics in Cognitive Science*, *6*(3), 371–389. https://doi.org/10.1111/tops.12102

Eck, E., Liu, J., Kazemzadeh-Atoufi, M., Ghoreishi, S., Blythe, S. A., & Garcia, H. G. (2020). Quantitative dissection of transcription in development yields evidence for transcription-factor-driven chromatin accessibility. *eLife*, *9*, e56429. https://doi.org/10.7554/eLife.56429

Elgin, C. (2007). Understanding and the Facts. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, *132*(1), 33–42. http://www.jstor.org/stable/25471843

Firth, J. R. (1957). A synopsis of linguistic theory 1930-55. In J. R. Firth, *Studies in Linguistic Analysis* (pp. 1–32). Blackwell.

Gettier, E. L. (1963). Is Justified True Belief Knowledge? *Analysis*, *23*(6), 121–123. https://doi.org/10.1093/analys/23.6.121

Haslanger, S. (2003). Gender and Race: (What) Are They? (What) Do We Want Them to Be? In A. Bird & J. Ladyman (Eds.), *Arguing About Science*, 95-116. Routledge.

Hempel, C. G., & Oppenheim, P. (1948). Studies in the Logic of Explanation. *Philosophy of Science*, *15*(2), 135–175. https://doi.org/10.1086/286983

Horák, A., & Rambousek, A. (2018). Lexicography and natural language processing. In P. A. Fuertes Olivera (Ed.), *The Routledge handbook of lexicography*. Routledge.

Isaac, A. M. C. (2019). The Semantics Latent in Shannon Information. *The British Journal for the Philosophy of Science*, *70*(1), 103–125. https://doi.org/10.1093/bjps/axx029

Jadacki, J. J. (1990). *O rozumieniu. Z filozoficznych podstaw semiotyki*. Wydawnictwa Uniwersytetu Warszawskiego.

Khalifa, K. (2017). *Understanding, explanation, and scientific knowledge*. Cambridge University Press.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., & Suchomel, V. (2014). The Sketch Engine: Ten years on. *Lexicography*, *1*(1), 7–36. https://doi.org/10.1007/s40607-014-0009-9

Kim, J. (1994). Explanatory Knowledge and Metaphysical Dependence. *Philosophical Issues*, *5*, 51. https://doi.org/10.2307/1522873

Koch S., Löhr G., & Pinder M. (2023). Recent work in the theory of conceptual engineering, *Analysis*, 83(3), 589–603. https://doi.org/10.1093/analys/anad032

Lean, O. M., Rivelli, L., & Pence, C. H. (2021). Digital Literature Analysis for Empirical Philosophy of Science. *The British Journal for the Philosophy of Science*, 715049. https://doi.org/10.1086/715049

Lenci, A. (2008). Distributional semantics in linguistic and cognitive research. *Italian Journal of Linguistics*, *20*(1), 1–31. https://www.italian-journal-linguistics.com/app/uploads/2021/05/1_Lenci.pdf

Lenci, A. (2018). Distributional Models of Word Meaning. *Annual Review of Linguistics*, *4*(1), 151–171. https://doi.org/10.1146/annurev-linguistics-030514-125254

Leslie, S.-J. (2017). The Original Sin of Cognition: Fear, Prejudice, and Generalization. *Journal of Philosophy*, *114*(8), 393–421. https://doi.org/10.5840/jphil2017114828

Machamer, P., Darden, L., & Craver, C. F. (2000). Thinking about Mechanisms. *Philosophy of Science*, *67*(1), 1–25. https://doi.org/10.1086/392759

Machery, E. (2004). Semantics, cross-cultural style. *Cognition*, *92*(3), B1–B12. https://doi.org/10.1016/j.cognition.2003.10.003

Malaterre, C., Chartier, J.-F., & Pulizzotto, D. (2019). What Is This Thing Called *Philosophy of Science* ? A Computational Topic-Modeling Perspective, 1934–2015. *HOPOS: The Journal of the International Society for the History of Philosophy of Science*, *9*(2), 215–249. https://doi.org/10.1086/704372

Mickus, T., Paperno, D., Constant, M., & van Deemter, K. (2020). What do you mean, BERT? Assessing BERT as a Distributional Semantics Model. *Proceedings of the Society for Computation in Linguistics 2020*, 279–290. https://doi.org/10.7275/T778-JA71

Miłkowski, M. (2022). Cognitive Artifacts and Their Virtues in Scientific Practice. *Studies in Logic, Grammar and Rhetoric*, *67*(3), 219–246. https://doi.org/10.2478/slgr-2022-0012

Miłkowski, M., & Jasieński, K. (2022). *eLife Open Peer Review Corpus* [dataset]. RepOD. https://doi.org/10.18150/FKPEQN

Miłkowski, M., Jasieński, K., & Depta, R. (2023). *MDPI Open Peer Review Corpus 2* [dataset]. RepOD. https://doi.org/10.18150/SHKP7B

Mitchell, M. (2019). *Artificial intelligence: A guide for thinking humans*. Farrar, Straus and Giroux.

Moretti, F. (2013). *Distant reading*. Verso.

Muñoz, M. M., Hu, Y., Anderson, P. S. L., & Patek, S. (2018). Strong biomechanical relationships bias the tempo and mode of morphological evolution. *eLife*, *7*, e37621. https://doi.org/10.7554/eLife.37621

Murdock, J., Allen, C., Börner, K., Light, R., McAlister, S., Ravenscroft, A., Rose, R., Rose, D., Otsuka, J., Bourget, D., Lawrence, J., & Reed, C. (2017). Multi-level computational methods for interdisciplinary research in the HathiTrust Digital Library. *PLOS ONE*, *12*(9), e0184188. https://doi.org/10.1371/journal.pone.0184188

Nandagopal, N., Santat, L. A., & Elowitz, M. B. (2019). Cis-activation in the Notch signaling pathway. *eLife*, *8*, e37880. https://doi.org/10.7554/eLife.37880

OpenAI. (2023, March 27). *GPT-4 Technical Report*. http://arxiv.org/abs/2303.08774

Overton, J. A. (2013). "Explain" in scientific discourse. *Synthese*, *190*(8), 1383–1405. https://doi.org/10.1007/s11229-012-0109-8

Páez, A. (2019). The Pragmatic Turn in Explainable Artificial Intelligence (XAI). *Minds and Machines*, *29*(3), 441–459. https://doi.org/10.1007/s11023-019-09502-w

Pichler, A., & Reiter, N. (2022). From Concepts to Texts and Back: Operationalization as a Core Activity of Digital Humanities. *Journal of Cultural Analytics*, *7*(4). https://doi.org/10.22148/001c.57195

Pitt, J. C. (2001). The Dilemma of Case Studies: Toward a Heraclitian Philosophy of Science. *Perspectives on Science*, *9*(4), 373–382. https://doi.org/10.1162/106361401760375785

Potochnik, A. (2017). *Idealization and the aims of science*. The University of Chicago Press.

Regt, H. W. de. (2017). *Understanding scientific understanding*. Oxford university press.

Rescher, N. (1979). *Cognitive systematization: A systems-theoretic approach to a coherentist theory of knowledge*. Basil Blackwell.

Rowbottom, D. P. (2014). Aimless science. *Synthese*, *191*(6), 1211–1221. https://doi.org/10.1007/s11229-013-0319-8

Sękowski, K. (2022). Concept Revision, Concept Application and the Role of Intuitions in Gettier Cases. *Episteme*, 1–19. https://doi.org/10.1017/epi.2022.49

Stich, S., & Tobia, K. P. (2016). Experimental Philosophy and the Philosophical Tradition. In J. Sytsma & W. Buckwalter (Eds.), *A Companion to Experimental Philosophy* (1st ed., pp. 3–21). Wiley. https://doi.org/10.1002/9781118661666.ch1

Strevens, M. (2013). No understanding without explanation. *Studies in History and Philosophy of Science Part A*, *44*(3), 510–515. https://doi.org/10.1016/j.shpsa.2012.12.005

Tanswell, F.S. (2018) Conceptual engineering for mathematical concepts, *Inquiry*, 61:8, 881-913, DOI: 10.1080/0020174X.2017.1385526

Turney, P. D. (2013). Distributional Semantics Beyond Words: Supervised Learning of Analogy and Paraphrase. *Transactions of the Association for Computational Linguistics*, *1*, 353–366. https://doi.org/10.1162/tacl_a_00233

Wilkenfeld, D. A. (2019). Understanding as compression. *Philosophical Studies*, *176*(10), 2807–2831. https://doi.org/10.1007/s11098-018-1152-1

Wittgenstein, L. (1968). *Philosophical investigations*. Basil Blackwell.

Yarkoni, T., & Westfall, J. (2017). Choosing Prediction Over Explanation in Psychology: Lessons From Machine Learning. *Perspectives on Psychological Science*, *12*(6), 1100–1122. https://doi.org/10.1177/1745691617693393